

1 The power and limitations of genomics to track 2 COVID-19 outbreaks: a case study from New Zealand

3 Jemma L Geoghegan PhD^{*1,2}, Jordan Douglas PhD^{*3}, Xiaoyun Ren PhD², Matthew Storey BSc²,
4 James Hadfield PhD⁴, Olin K Silander PhD⁵, Nikki E Freed PhD⁵, Lauren Jelley MAppSc², Sarah
5 Jefferies MD², Jillian Sherwood MBChB², Shevaun Paine MAE², Sue Huang PhD², Andrew Sporle
6 MA^{6,7}, Michael G Baker MBChB⁸, David R Murdoch MD⁹, Alexei J Drummond PhD³, David Welch
7 PhD³, Colin R Simpson PhD^{10,11}, Nigel French PhD¹², Edward C Holmes PhD¹³, Joep de Ligt PhD².

8

9 ¹Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand.

10 ²Institute of Environmental Science and Research, Wellington, New Zealand.

11 ³Centre for Computational Evolution, School of Computer Science, University of Auckland,
12 Auckland, New Zealand.

13 ⁴Fred Hutchinson Cancer Research Centre, Seattle, Washington, USA.

14 ⁵School of Natural and Computational Sciences, Massey University, Auckland, New Zealand.

15 ⁶Department of Statistics, University of Auckland, New Zealand.

16 ⁷iNZight Analytics Ltd., Auckland, New Zealand.

17 ⁸Department of Public Health, University of Otago, Wellington, New Zealand.

18 ⁹Department of Pathology and Biomedical Science, University of Otago, Christchurch, New
19 Zealand.

20 ¹⁰School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, New
21 Zealand.

22 ¹¹Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.

23 ¹²School of Veterinary Science, Massey University, Palmerston North, New Zealand.

24 ¹³Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental
25 Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales,
26 Australia.

27 *contributed equally

28 Author for correspondence: jemma.geoghegan@otago.ac.nz

29 **NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**
Keywords: SARS-CoV-2; COVID-19; coronavirus; genomics; infectious disease; New Zealand

30 Summary

31 **Background.** Real-time genomic sequencing has played a major role in tracking the global
32 spread and local transmission of SARS-CoV-2, contributing greatly to disease mitigation
33 strategies. After effectively eliminating the virus, New Zealand experienced a second outbreak of
34 SARS-CoV-2 in August 2020. During this August outbreak, New Zealand utilised genomic
35 sequencing in a primary role to support its track and trace efforts for the first time, leading to a
36 second successful elimination of the virus.

37 **Methods.** We generated the genomes of 80% of the laboratory-confirmed samples of SARS-
38 CoV-2 from New Zealand's August 2020 outbreak and compared these genomes to the available
39 global genomic data.

40 **Findings.** Genomic sequencing was able to rapidly identify that the new COVID-19 cases in New
41 Zealand belonged to a single cluster and hence resulted from a single introduction. However,
42 successful identification of the origin of this outbreak was impeded by substantial biases and
43 gaps in global sequencing data.

44 **Interpretation.** Access to a broader and more heterogenous sample of global genomic data
45 would strengthen efforts to locate the source of any new outbreaks.

46 **Funding.** This work was funded by the Ministry of Health of New Zealand, New Zealand Ministry
47 of Business, Innovation and Employment COVID-19 Innovation Acceleration Fund (CIAF-0470),
48 ESR Strategic Innovation Fund and the New Zealand Health Research Council (20/1018 and
49 20/1041).

50

51 Main Text

52 Only twelve days after the novel coronavirus SARS-CoV-2 was identified, a genome of the virus
53 was first published¹. This information was pivotal to the subsequent, rapid, development of
54 diagnostic tests and identification of potential treatments^{2,3}. Since then, as of 25 September 2020,
55 over 110,000 genomes of SARS-CoV-2 have been shared publicly⁴. The underlying genome
56 sequencing has occurred so quickly that, for the first time during an infectious disease outbreak,
57 it has enabled virological and epidemiological data to be integrated in real time⁵. Analysis of these
58 data has played an important role in informing the COVID-19 response by tracking the global
59 spread and evolution of SARS-CoV-2, including identification of the number, source, and timing
60 of introductions into individual countries. This has led to a greater understanding of COVID-19
61 outbreaks around the world⁶⁻¹³.

62 Of the 185 countries that have reported positive cases of COVID-19 to the World Health
63 Organization¹⁴, 60% (n=112) have sequenced and shared SARS-CoV-2 genomes on the GISAID
64 database⁴ (September 2020). This immense global sequencing effort has facilitated ongoing

65 genomic surveillance of the pandemic, including monitoring viral genetic changes of interest¹²,
66 which has informed public health responses¹⁵⁻¹⁸. Nevertheless, the number and proportion of
67 positive cases sequenced, and genomes published, varies dramatically between countries and
68 over time (Figure 1). The COVID-19 Genomics UK (COG-UK) consortium, for example, has led to
69 the UK being the most represented sampling location, totalling over 42,000 genomes and
70 comprising 39% of the global data set despite recording only 1% of the world's positive cases (n
71 = 412,241). Conversely, SARS-CoV-2 genomes sequenced in India represent just 3% of the
72 global data set but 18% of the world's total reported cases (n = 5,646,010).

73 We show here that such disparities in sequencing effort can have major implications for data
74 interpretation and must be met with careful consideration. Real-time sequencing of SARS-CoV-2
75 genomes has, however, had particular utility in tracking the re-emergence of the virus in New
76 Zealand. By June 2020 New Zealand had effectively eliminated COVID-19 in the community and
77 positive cases were limited to those linked to managed quarantine facilities at the border^{7,13,19}.
78 Following over 100 days with no detected community transmission of COVID-19, four new cases
79 emerged on August 12, 2020 with no apparent epidemiological link to any known case. We
80 compared the genomes of these cases to sequenced cases from both New Zealand's first wave
81 and those in quarantine facilities and, again, found no link. The vast majority of available
82 sequence data from cases in New Zealand's quarantine facilities were of different virus lineages
83 than that of the August 2020 outbreak. However, this observation was of limited value given that
84 only 42% of cases in those quarantine facilities had adequate viral RNA for successful genomic
85 sequencing. In order to determine the likely origins of this outbreak we compared genomes from
86 the new community outbreak to the global data set.

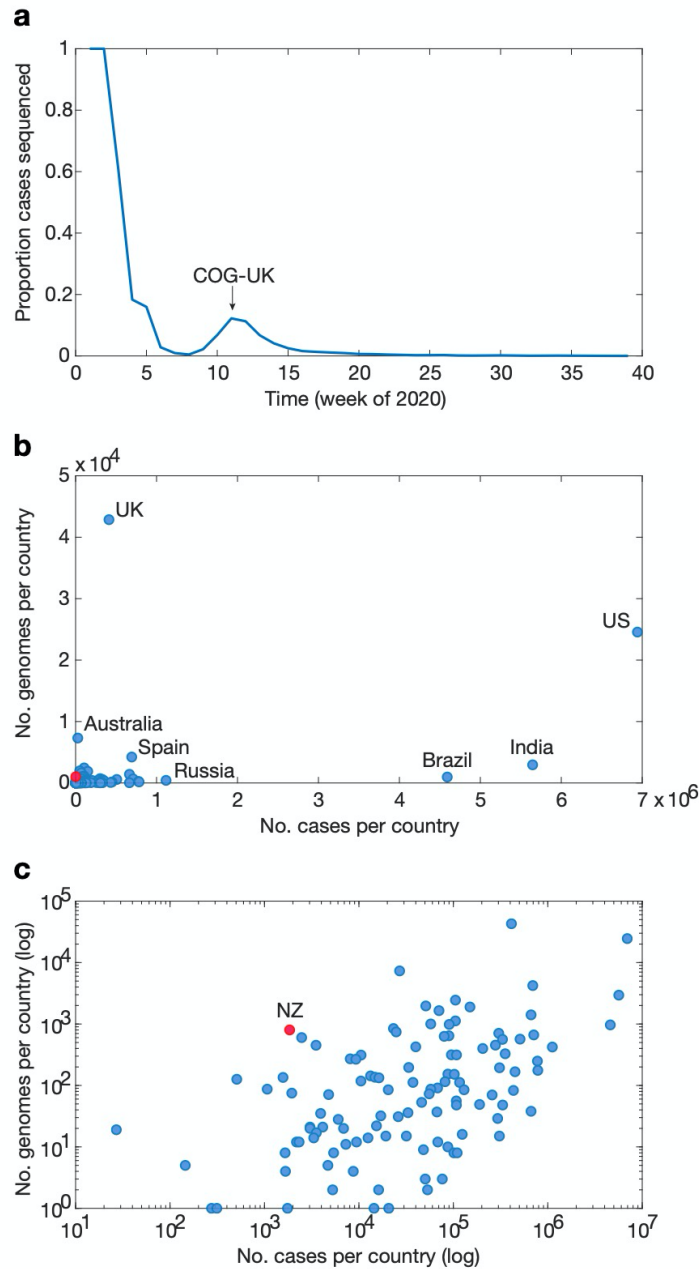
87 An initial genomic sequence analysis found that the re-emergence of COVID-19 in New Zealand
88 was due to a SARS-CoV-2 from lineage B.1.1.1²⁰. Of the countries that have contributed SARS-
89 CoV-2 data, 40% had genomes of this lineage. Remarkably, 85% of B.1.1.1. genomes were from
90 the UK and were generated between March – September 2020, the vast majority of which were
91 sampled during the first wave of disease in the UK (Figure 2). Phylogenetic analysis of the most
92 recently sampled B.1.1.1. genomes (1,996 of 4,544) identified genomes sampled from
93 Switzerland, South Africa and England in August as the most likely to be contained within the
94 sister clade (Figure 2): these were the closest sampled genomic relatives of the viruses associated
95 with New Zealand's August 2020 outbreak (Supplementary Figure 1). Additional Bayesian analysis
96 estimated that the outbreak originated 11 days before the first transmission event, with a 95%
97 highest posterior density (HPD) of 0 – 28 days. We also estimated that the first transmission event
98 in the outbreak occurred between 26 July – 13 August 2020 (95% HPD; mean date of 5 August).
99 Epidemiological data showed that two confirmed cases linked to the outbreak had a symptom
100 onset date of 31 July, although the most probable sampled genomes within the sister clade were

101 sampled later, between 6 August – 28 August. Hence, is unlikely that the currently available global
102 genomic data set contains the source of this outbreak.

103 Genomic epidemiological analysis on the possible origins of New Zealand’s re-emergence was
104 found to be inconclusive, likely due to missing genomic data within the quarantine border facilities
105 as well as in the global data set. A glimpse into the genomic diversity likely omitted from the
106 global data set can be seen in the genomes sequenced in New Zealand from positive quarantine
107 cases which comprise citizens and residents returning from across the globe . For example, 12
108 SARS-CoV-2 genomes from returnees to New Zealand from India who arrived on the same flight
109 fell across at least four genomic lineages and comprised sequence divergence of up to 34 single
110 nucleotide polymorphisms. This represented far more genomic mutations than was observed in
111 New Zealand during the first outbreak in March – May 2020. Such a high level of diversity in just a
112 small sample of positive cases from India suggest that the currently available genomic data fails
113 to encompass the true diversity that existed locally let alone globally.

114 The re-emergence of COVID-19 into New Zealand in August 2020 exemplified one of the most
115 complete genomic data sets for a specific outbreak compiled to date, comprising 81% of positive
116 cases (145 of 179 total PCR positive cases). Real-time genomic sequencing quickly informed
117 track and trace efforts to control the outbreak, setting New Zealand on track to eliminate the virus
118 from the community for the second time. The rapid genome sequencing of positive samples
119 provided confidence to public health teams regarding links to the outbreak and identified that
120 cases and sub-clusters were linked to a single genomic lineage, resulting from a single
121 introduction event. Indeed, the timing and length of lockdown measures were partly informed
122 based on these data.

123 Nevertheless, the biased nature of global sampling, including the contribution of very few genome
124 sequences from certain geographic locations, clearly limited the power of genomics to attribute
125 the geographical origin of New Zealand’s August 2020 outbreak. We therefore advocate that
126 careful consideration of the potential sampling biases and gaps in available genomic data be
127 made whenever attempting to determine the geographic origins of a specific outbreak of SARS-
128 CoV-2. Analysis should consider all available evidence, including from genomic and
129 epidemiological sources.



130

131 **Figure 1. a** Proportion of global cases sequenced and shared on GISAID between December
132 2019 to September 2020, where the second mode was largely driven by the COG-UK consortium
133 as illustrated; **b** Number of genomes sequenced and number of reported cases per country (New
134 Zealand is represented in red) on a linear scale; **c** Number of genomes sequenced and number of
135 reported cases per country on a logarithmic scale.

142 Methods

143 **Ethics statement.** Nasopharyngeal samples testing positive for SARS-CoV-2 by real-time
144 polymerase chain reaction (RT-PCR) were obtained from public health medical diagnostics
145 laboratories located throughout New Zealand. All samples were de-identified before receipt by the
146 researchers. Under contract for the New Zealand Ministry of Health, ESR has approval to conduct
147 genomic sequencing for surveillance of notifiable diseases.

148 **Genomic sequencing of SARS-CoV-2: New Zealand's second wave.** A total of 172 (out of 179)
149 laboratory-confirmed samples of SARS-CoV-2 were received by ESR for whole genome
150 sequencing, comprising samples from New Zealand's August 2020 outbreak. Genome
151 sequencing of SARS-CoV-2 samples was performed as before⁷. In short, viral extracts were
152 prepared from respiratory tract samples where SARS-CoV-2 was detected by RT-PCR using
153 WHO recommended primers and probes targeting the E and N gene. Extracted RNA from SARS-
154 CoV-2 positive samples were subject to whole genome sequencing following the ARTIC network
155 protocol (V3) (<https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye>)
156 and the Massey University 1200 bp primer set (<https://www.protocols.io/view/ncov-2019-sequencing-protocol-rapid-barcoding-1200-bh7hj9j6>)²¹.

158 Briefly, one of the tiling amplicon designs were used to amplify viral cDNA prepared with
159 SuperScript IV. Sequence libraries were then constructed, using the Oxford Nanopore ligation
160 sequencing and native barcoding expansion kit for samples amplified with the ARTIC V3 primer
161 sets, and the Oxford Nanopore rapid barcoding kit for samples amplified with the 1200 bp primer
162 sets. The 1200 bp primers and rapid barcoding was used in cases where the genomes were
163 required urgently. Libraries were sequenced using R9.4.1 MinION flow cells, respectively. Near-
164 complete (>90% recovered) viral genomes were subsequently assembled through reference
165 mapping. Steps included in the pipeline are described in detail online ([https://github.com/ESR-
166 NZ/NZ SARS-CoV-2_genomics](https://github.com/ESR-NZ/NZ_SARS-CoV-2_genomics)). The reads generated with Nanopore sequencing using ARTIC
167 primer sets (V3) were mapped and assembled using the ARTIC bioinformatics medaka pipeline (v
168 1.1.0). In total, 145 (out of 172) genomes from New Zealand's August 2020 outbreak passed
169 quality control. All data are available on GISAID.

170 **Phylogenetic analysis of SARS-CoV-2.** All human SARS-CoV-2 genomes available on GISAID⁴
171 as of 25 September 2020 (n = 109,379) were downloaded and their lineages assigned according
172 to the proposed nomenclature¹⁶ using pangolin (<https://github.com/hCoV-2019/pangolin>).
173 Genomes assigned to the B.1.1.1. lineage (n = 4,544) were subsampled to include 1,996 most
174 recent-in-time sequences along with four outgroup (non-B.1.1.1.) sequences, and were aligned
175 with those from the recent New Zealand outbreak (n = 140) using MAFFT (v 7)²² employing the
176 FFT-NS-2 progressive alignment algorithm. Bayesian phylogenetic analyses were performed

177 using BEAST 2.5²³. We used a strict clock model with one HKY substitution model (estimated
178 frequencies) for each codon position and one for non-coding positions. We employed the
179 Bayesian skyline model²⁴ as a tree prior to allow effective population sizes to change over time
180 intervals. These components of the model and their prior distributions are those used by Douglas
181 et al. 2020¹³. Phylogenetic trees were annotated using FigTree (v 1.4)²⁵ and Tree of Life (v 4)²⁶.

182 Acknowledgements

183 We thank the ARTIC network for making their protocols and tools openly available and specifically
184 Josh Quick for sending the initial V1 and V3 amplification primers. We thank the diagnostic
185 laboratories that performed the initial RT-PCRs and referred samples for sequencing as well as
186 the public health units for providing epidemiological data. We thank Genomics Aotearoa for their
187 support, the NextStrain team for their support and timely global and local analysis, and all those
188 who have contributed SARS-CoV-2 sequences to GenBank and GISAID databases. The authors
189 wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance
190 computing facilities. New Zealand's national facilities are provided by NeSI and funded jointly by
191 NeSI's collaborator institutions and through the Ministry of Business, Innovation and
192 Employment's Research Infrastructure programme (<https://www.nesi.org.nz>).

193 References

- 194 1. Holmes EC. Novel 2019 coronavirus genome. 2020. [https://virological.org/t/novel-2019-
195 coronavirus-genome/319](https://virological.org/t/novel-2019-
195 coronavirus-genome/319).
- 196 2. Shin MD, Shukla S, Chung YH, et al. COVID-19 vaccine development and a potential
197 nanomaterial path forward. *Nature Nanotechnology* 2020; **15**(8): 646-55.
- 198 3. Stebbing J, Phelan A, Griffin I, et al. COVID-19: combining antiviral and anti-inflammatory
199 treatments. *Lancet Infect Dis* 2020; **20**(4): 400-2.
- 200 4. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative
201 contribution to global health. *Global Challenges* 2017; **1**(1): 33-46.
- 202 5. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution.
203 *Bioinformatics* 2018; **34**(23): 4121-3.
- 204 6. Bedford T, Greninger AL, Roychoudhury P, et al. Cryptic transmission of SARS-CoV-2 in
205 Washington State. *medRxiv* 2020: 2020.04.02.20051417.
- 206 7. Geoghegan JL, Ren X, Storey M, et al. Genomic epidemiology reveals transmission
207 patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *medRxiv* 2020:
208 2020.08.05.20168930.
- 209 8. Candido DS, Claro IM, de Jesus JG, et al. Evolution and epidemic spread of SARS-CoV-2
210 in Brazil. *Science* 2020: eabd2161.
- 211 9. Eden J-S, Rockett R, Carter I, et al. An emergent clade of SARS-CoV-2 linked to returned
212 travellers from Iran. *Virus Evolution* 2020; **6**(1).
- 213 10. Filipe ADS, Shepherd J, Williams T, et al. Genomic epidemiology of SARS-CoV-2 spread in
214 Scotland highlights the role of European travel in COVID-19 emergence. *medRxiv* 2020:
215 2020.06.08.20124834.
- 216 11. Seemann T, Lane C, Sherry N, et al. Tracking the COVID-19 pandemic in Australia using
217 genomics. *medRxiv* 2020: 2020.05.12.20099929.
- 218 12. Zhang L, Jackson CB, Mou H, et al. The D614G mutation in the SARS-CoV-2 spike protein
219 reduces S1 shedding and increases infectivity. *bioRxiv* 2020: 2020.06.12.148726.
- 220 13. Douglas J, Mendes FK, Bouckaert R, et al. Phylodynamics reveals the role of human travel
221 and contact tracing in controlling COVID-19 in four island nations. *medRxiv* 2020:
222 2020.08.04.20168518.
- 223 14. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real
224 time. *The Lancet Infectious Diseases* 2020; **20**(5): 533-4.
- 225 15. Oude Munnink BB, Nieuwenhuijse DF, Stein M, et al. Rapid SARS-CoV-2 whole-genome
226 sequencing and analysis for informed public health decision-making in the Netherlands. *Nature*
227 *Medicine* 2020; **26**(9): 1405-10.
- 228 16. Kalinich CC, Jensen CG, Neugebauer P, et al. Real-time public health communication of
229 local SARS-CoV-2 genomic epidemiology. *PLOS Biology* 2020; **18**(8): e3000869.
- 230 17. Rockett RJ, Arnott A, Lam C, et al. Revealing COVID-19 transmission in Australia by
231 SARS-CoV-2 genome sequencing and agent-based modeling. *Nature Medicine* 2020; **26**(9):
232 1398-404.
- 233 18. Bauer DC, Tay AP, Wilson LOW, et al. Supporting pandemic response using genomics
234 and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. *Transboundary and*
235 *Emerging Diseases* 2020; **67**(4): 1453-62.
- 236 19. Jefferies S, French N, Gilkison C, et al. COVID-19 in New Zealand and the impact of the
237 national response: a descriptive epidemiological study. *The Lancet Public Health*.
- 238 20. Rambaut A, Holmes EC, Hill V, et al. A dynamic nomenclature proposal for SARS-CoV-2
239 to assist genomic epidemiology. *bioRxiv* 2020: 2020.04.17.046086.
- 240 21. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome
241 sequencing of SARS-CoV-2 using 1200bp tiled amplicons and Oxford Nanopore Rapid
242 Barcoding. *Biology Methods and Protocols* 2020; **5**(1).
- 243 22. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
244 improvements in performance and usability. *Mol Biol Evol* 2013; **30**(4): 772-80.

- 245 23. Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: An advanced software
246 platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 2019; **15**(4): e1006650.
247 24. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past
248 population dynamics from molecular sequences. *Mol Biol Evol* 2005; **22**(5): 1185-92.
249 25. Rambaut A, Drummond AJ. FigTree version 1.4.0. 2012.
250 26. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new
251 developments. *Nucleic Acids Research* 2019; **47**(W1): W256-W9.
252